Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen[1]

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

The ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age in which numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization, and quantification of sequence similarity between the gene of interest and genes for which functional information is available. Because sequence is the prime determining factor of function, sequence similarity is taken to imply similarity of function. There is no doubt that this assumption is valid in most cases. However, sequence similarity does not ensure identical functions, and it is common for groups of genes that are similar in sequence to have diverse (although usually related) functions. Therefore, the identification of sequence similarity is frequently not enough to assign a predicted function to an uncharacterized gene; one must have a method of choosing among similar genes with different functions. In such cases, most functional prediction methods assign likely functions by quantifying the levels of similarity among genes. I suggest that functional predictions can be greatly improved by focusing on *how* the genes became similar in sequence (i.e., evolution) rather than on the sequence similarity itself. It is well established that many aspects of comparative biology can benefit from evolutionary studies (Felsenstein 1985), and comparative molecular biology is no exception (e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic* approach to the prediction of gene function, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

[1]E-MAIL jeisen@leland.stanford.edu; FAX (650) 725-1848.
WWW: http://www-leland.stanford.edu/~jeisen.

Sequence Similarity, Homology, and Functional Predictions

To make use of the identification of sequence similarity between genes, it is helpful to understand how such similarity arises. Genes can become similar in sequence either as a result of *convergence* (similarities that have arisen without a common evolutionary history) or descent with modification from a common ancestor (also known as *homology*). It is imperative to recognize that sequence similarity and homology are not interchangeable terms. Not all homologs are similar in sequence (i.e., homologous genes can diverge so much that similarities are difficult or impossible to detect) and not all similarities are due to homology (Reeck et al. 1987; Hillis 1994). Similarity due to convergence, which is likely limited to small regions of genes, can be useful for some functional predictions (Henikoff et al. 1997). However, most sequence-based functional predictions are based on the identification (and subsequent analysis) of similarities that are thought to be due to homology. Because homology is a statement about common ancestry, it cannot be proven directly from sequence similarity. In these cases, the inference of homology is made based on finding levels of sequence similarity that are thought to be too high to be due to convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions because not all homologs have the same function. The available similarity-based functional prediction methods can be distinguished by how they choose the homolog whose function is most relevant to a particular uncharacterized gene (Table 1). Some methods are relatively simple—many researchers use the highest scoring homolog (as determined by programs like BLAST or BLAZE) as the basis for assigning function. While highest hit methods are very fast, can be automated readily, and are likely accurate in many instances, they do not take advantage of any information about how genes and gene functions evolve. For example, gene duplication and subsequent divergence of function of the duplicates can result in homologs with different functions being present within one species. Specific terms have been created to distinguish homologs in these cases (Table 2): Genes of the same duplicate group are called *orthologs* (e.g., β-globin from mouse and humans), and different duplicates are called *paralogs* (e.g., α- and β-globin) (Fitch 1970). Because gene duplications are frequently accompanied by functional divergence, dividing genes into groups of orthologs and paralogs can improve the accuracy of functional predictions. Recognizing that the one-to-one sequence comparisons used by most methods do not reliably distinguish orthologs from paralogs, Tatusov et al. (1997) developed the COG cluster-

8:163–167 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

Table 1. Methods of Predicting Gene Function When Homologs Have Multiple Functions

Highest Hit	The uncharacterized gene is assigned the function (or frequently, the annotated function) of the gene that is identified as the highest hit by a similarity search program (e.g., Tomb et al. 1997).
Top Hits	Identify top 10+ hits for the uncharacterized gene. Depending on the degree of consensus of the functions of the top hits, the query sequence is assigned a specific function, a general activity with unknown specificity, or no function (e.g., Blattner et al. 1997).
Clusters of Orthologous Groups	Genes are divided into groups of orthologs based on a cluster analysis of pairwise similarity scores between genes from different species. Uncharacterized genes are assigned the function of characterized orthologs (Tatusov et al. 1997).
Phylogenomics	Known functions are overlaid onto an evolutionary tree of all homologs. Functions of uncharacterized genes are predicted by their phylogenetic position relative to characterized genes (e.g., Eisen et al. 1995, 1997).

ing method (see Table 1). Although the COG method is clearly a major advance in identifying orthologous groups of genes, it is limited in its power because clustering is a way of classifying levels of similarity and is not an accurate method of inferring evolutionary relationships (Swofford et al. 1996). Thus, as sequence similarity and clustering are not reliable estimators of evolutionary relatedness, and as the incorporation of such phylogenetic information has been so useful to other areas of biology, evolutionary techniques should be useful for improving the accuracy of predicting function based on sequence similarity.

Phylogenomics

There are many ways in which evolutionary information can be used to improve functional predictions. Below, I present an outline of one such *phylogenomic* method (see Fig. 1), and I compare this method to nonevolutionary functional prediction methods. This method is based on a relatively simple assumption—because gene functions change as a result of evolution, reconstructing the evolutionary history of genes should help predict the functions of uncharacterized genes. The first step is the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. Such trees are distinct from clusters and other means of characterizing sequence similarity because they are inferred by special techniques that help convert patterns of similarity into evolutionary relationships (see Swofford et al. 1996). After the gene tree is inferred, biologically determined functions of the various homologs are overlaid onto the tree. Finally, the structure of the tree and the relative phylogenetic positions of genes of different functions are used to trace the history of functional changes, which is then used to predict functions of uncharacterized genes. More detail of this method is provided below.

Identification of Homologs

The first step in studying the evolution of a particular gene is the identification of homologs. As with similarity-based functional prediction methods, likely homologs of a particular gene are identified through database searches. Because phylogenetic methods benefit greatly from more data, it is useful to augment this initial list by using identified homologs as queries for further database searches or using automatic iterated search methods such as PSI-BLAST (Altschul et al. 1997). If a gene family is very large (e.g., ABC transporters), it may be necessary to only analyze a subset of homologs. However, this must be done with extreme care, as one might accidentally leave out proteins that would be important for the analysis.

Alignment and Masking

Sequence alignment for phylogenetic analysis has a particular purpose—it is the assignment of *positional* homology. Each column in a multiple sequence alignment is assumed to include amino acids or nucleotides that have a common evolutionary history, and each column is treated separately in the phylogenetic analysis. Therefore, regions in which the assignment of positional homology is ambiguous should be excluded (Gatesy et al. 1993). The exclusion of certain alignment positions (also known as masking) helps to give phylogenetic methods much of their discriminatory power. Phylogenetic trees generated without masking (as is done in many sequence analysis software packages) are less likely to accurately reflect the evolution of the genes than trees with masking.

Phylogenetic Trees

For extensive information about generating phylogenetic trees from sequence alignments, see Swofford et al. (1996). In summary, there are three methods commonly used: parsimony, distance, and maximum likelihood (Table 3), and each has its advantages and disadvantages. I

Table 2. Types of Molecular Homology

Homolog	Genes that are descended from a common ancestor (e.g., all globins)
Ortholog	Homologous genes that have diverged from each other after *speciation* events (e.g., human β- and chimp β-globin)
Paralog	Homologous genes that have diverged from each other after *gene duplication* events (e.g., β- and γ-globin)
Xenolog	Homologous genes that have diverged from each other after *lateral gene transfer* events (e.g., antibiotic resistance genes in bacteria)
Positional homology	Common ancestry of specific amino acid or nucleotide positions in different genes

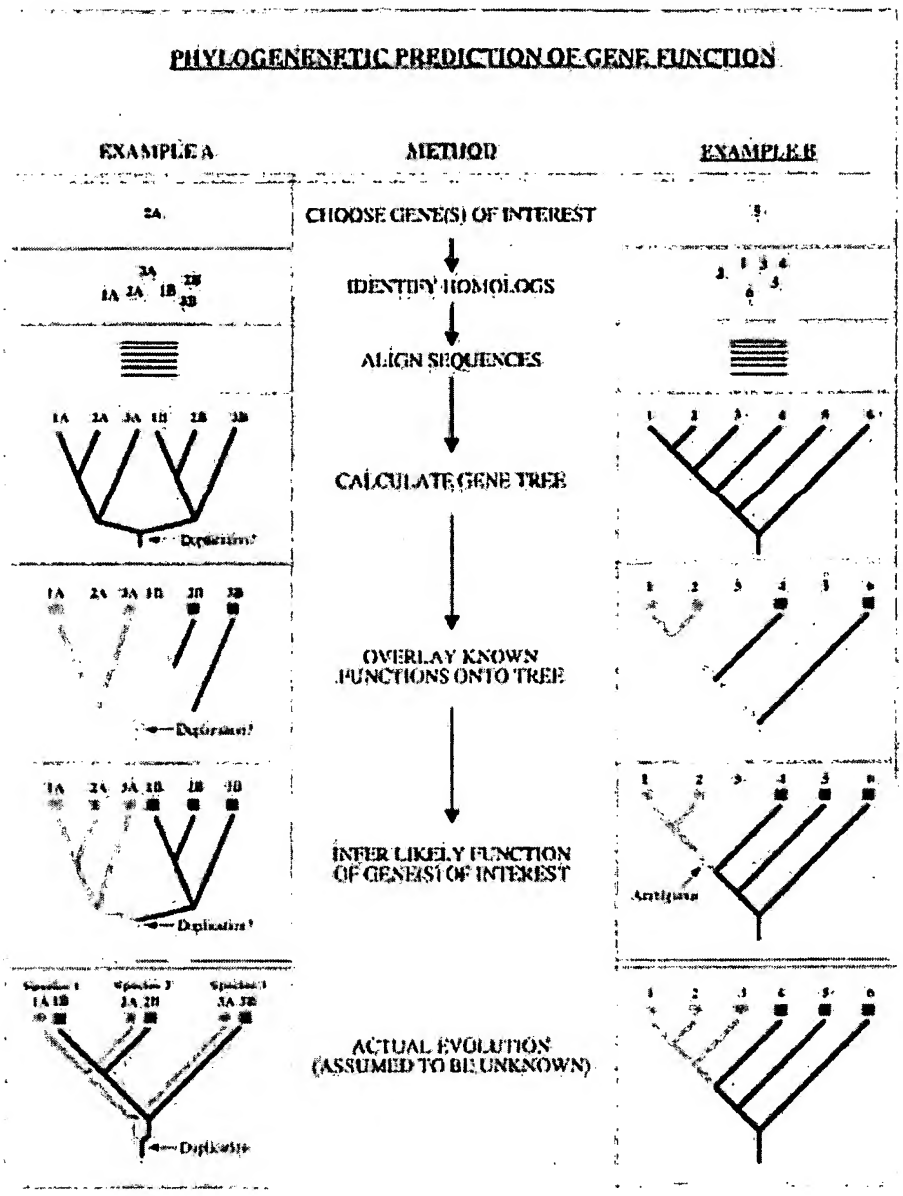

Figure 1 Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (*A*) A gene family has undergone a gene duplication that was accompanied by functional divergence. (*B*) Gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the *bottom*. The genes are referred to by numbers (which represent the species from which these genes come) and letters (which in *A* represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in *A* (*bottom*) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions; gray (with hatching) represents either unknown or unpredictable functions.

prefer distance methods because they are the quickest when using large data sets. Before using any particular tree it is important to estimate the robustness and accuracy of the phylogenetic patterns it shows (through techniques such as the comparison of trees generated by different methods and bootstrapping). Finally, in most cases, it is also useful to determine a root for the tree.

Functional Predictions

To make functional predictions based on the phylogenetic tree, it is necessary to first overlay any known functions onto the tree. There are many ways this "map" can then be used to make functional predictions, but I recommend splitting the task into two steps. First, the tree can be used to identify likely gene duplication events in the past. This allows the division of the genes into groups of orthologs and paralogs (e.g., Eisen et al. 1995). Uncharacterized genes can be assigned a likely function if the function of any ortholog is known (and if all characterized orthologs have the same function). Second, parsimony reconstruction techniques (Maddison and Maddison 1992) can be used to infer the likely functions of uncharacterized genes by identifying the evolutionary scenario that requires the fewest functional changes over time (Fig. 1). The incorporation of more realistic models of functional change (and not just minimizing the total number of changes) may prove to be useful, but the parsimony minimization methods are probably sufficient in most cases.

Is the Phylogenomic Method Worth the Trouble?

Phylogenomic methods require many more steps and usually much more manual labor than similarity-based functional prediction methods. Is the phylogenomic approach worth the trouble? Many specific examples exist in which gene function has been shown to correlate well with gene phylogeny (Eisen et al. 1995; Atchley and Fitch 1997). Although no systematic comparisons of phylogenetic versus similarity-based functional prediction methods have been done, there are a variety of reasons to believe that the phylogenomic method should produce more accurate predictions than similarity-based methods. In particular, there are many conditions in which similarity-based methods are likely to make inaccurate predictions but which can be dealt with well by phylogenetic methods (see Table 4).

A specific example helps illustrate a potential problem with similarity-based methods. Molecular phylogenetic methods show conclusively that mycoplasmas share a common ancestor with low-GC Gram-positive bacteria (Weisburg et

Table 3. Molecular Phylogenetic Methods

Method	
Parsimony	Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino acid substitutions) that would be required over evolutionary time to fit the sequences into that tree. The optimal tree is considered to be the one requiring the fewest changes (the most parsimonious tree).
Distance	The optimal tree is generated by first calculating the estimated evolutionary distance between all pairs of sequences. Then these distances are used to generate a tree in which the branch patterns and lengths best represent the distance matrix.
Maximum likelihood	Maximum likelihood is similar to parsimony methods in that possible trees are compared and given a score. The score is based on how likely the given sequences are to have evolved in a particular tree given a model of amino acid or nucleotide substitution probabilities. The optimal tree is considered to be the one that has the highest probability.
Bootstrapping	Alignment positions within the original multiple sequence alignment are resampled and new data sets are made. Each bootstrapped data set is used to generate a separate phylogenetic tree and the trees are compared. Each node of the tree can be given a bootstrap percentage indicating how frequently those species joined by that node group together in different trees. Bootstrap percentage does not correspond directly to a confidence limit.

al. 1989). However, examination of the percent similarity between mycoplasmal genes and their homologs in bacteria does not clearly show this relationship. This is because mycoplasmas have undergone an accelerated rate of molecular evolution relative to other bacteria. Thus, a BLAST search with a gene from *Bacillus subtilis* (a low GC Gram-positive species) will result in a list in which the mycoplasma homologs (if they exist) score lower than genes from many spe-

Table 4 Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function

Evolutionary Pattern and Typical Genes and Functions	Gene Web Example in Figure 1	Highest BLAST Method: Predicted Function	Highest BLAST Method: Accurate?	Phylogenomic Method: Predicted Function	Phylogenomic Method: Accurate?	Comments
A. Functional change during evolution	1		+		+	• Phylogenomic method cannot predict functions for all genes, but the predictions that are made are correct
	2		+		+	• Highest hit method is misleading because functions changed among homologs but firmer hits of similarity do not correlate with the functions (see Bork and Koonin 1996)
	3		−	■	±	
	4■		−	■	±	
	5■	■	±	■	+	
	6■	■	±	■	+	
B. Functional change & rate variation	1		+		+	• Similarity-based methods perform particularly poorly when evolutionary rates vary between taxa
	2		+		+	• Molecular phylogenetic methods can allow for rate variation and reconstruct gene history reasonably accurately
	3	■	−	■	±	
	4■		−	■	±	
	5■		−	■	±	
	6■	■	+	■	+	
C. Gene duplications and rate variation	1A		+		+	• Most similarity-based methods are not ideally set up to deal with cases of gene duplication since orthologous genes do not always have significantly more sequence similarity to each other than to paralogs (Eisen et al. 1998; Zdobnov et al. 1998; Thomas et al. 1997)
	2A		+		+	
	3A■	■	−		+	• Similarity-based methods perform particularly poorly when rate variation and gene duplication are combined. This even applies to the COG method (see Table 1) since it works by classifying levels of similarity and not by reconstructing history. Nevertheless, the COG method is a significant improvement over other similarity-based methods in classifying orthologs
	1B■	■	+	■	+	
	2B■	■	+	■	+	
	3B■		−	■	+	• Phylogenetic reconstruction is the most reliable way to infer gene duplication events and thus determine orthology

cies of bacteria less closely related to *B. subtilis*. When amounts or rates of change vary between lineages, phylogenetic methods are better able to infer evolutionary relationships than similarity methods (including clustering) because they allow for evolutionary branches to have different lengths. Thus, in those cases in which gene function correlates with gene phylogeny and in which amounts or rates of change vary between lineages, similarity-based methods will be more likely than phylogenomic methods to make inaccurate functional predictions (see Table 4).

Another major advantage of phylogenetic methods over most similarity methods comes from the process of masking (see above). For example, a deletion of a large section of a gene in one species will greatly affect similarity measures but may not affect the function of that gene. A phylogenetic analysis including these genes could exclude the region of the deletion from the analysis by masking. In addition, regions of genes that are highly variable between species are more likely to undergo convergence and such regions can be excluded from phylogenetic analysis by masking. Masking thus allows the exclusion of regions of genes in which sequence similarity is likely to be "noisy" or misleading rather than a biologically important signal. The pairwise sequence comparisons used by most similarity-based functional prediction methods do not allow such masking. Phylogenetic methods have been criticized because of their dependence (for most methods) on multiple sequence alignments that are not always reliable and unbiased. However, multiple sequence alignments also allow for masking, which is probably more valuable than the cost of depending on alignments.

The conditions described above and highlighted in Table 4 are just some examples of conditions in which evolutionary methods are more likely to make accurate functional predictions than similarity-based methods. Phylogenetic methods are particularly useful when the history of a gene family includes many of these conditions (e.g., multiple gene duplications plus rate variation) or when the gene family is very large. The principle is simple—the more complicated the history of a gene family, the more useful it is to try to infer that history. Thus although the phylogenomic method is slow and labor intensive, I believe it is worth using if accuracy is the main objective. In addition, information about the evolutionary relationships among gene homologs is useful for summarizing relationships among genes and for putting functional information into a useful context.

Despite the evolution of these methods, and likely continued improvements in functional predictions, it must be remembered that the key word is *prediction*. All methods are going to make inaccurate predictions of functions. For example, none of the methods described can perform well when gene functions can change with little sequence change as has been seen in proteins like opsins (Yokoyama 1997). Thus, sequence databases and genome researchers should make clear which functions assigned to genes are based on predictions and which are based on experiments. In addition, all prediction methods should use only experimentally determined functions as their grist for predictions. This will hopefully limit error propagation that can happen by using an inaccurate prediction of function to then predict the function of a new gene, which is a particular problem for the highest hit methods, as they rely on the function of only one gene at a time to make predictions (Eisen et al. 1997). Despite these and other potential problems, functional predictions are of great value in guiding research and in sorting through huge amounts of data. I believe that the increased use of phylogenetic methods can only serve to improve the accuracy of such functional predictions.

REFERENCES

Altschul, S.F., R.J. Carroll, and D.J. Lipman. 1989. *J. Mol. Biol.* **207:** 647–653.

Altschul, S.F., T.L. Madden, A.A. Schaeffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. *Nucleic Acids Res.* **25:** 3389–3402.

Atchley, W.R. and W.M. Fitch. 1997. *Proc. Natl. Acad. Sci.* **94:** 5172–5176.

Blattner, F.R., G.I. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. *Science* **277:** 1453–1462.

Bolker, J.A. and R.A. Raff. 1996. *BioEssays* **18:** 489–494.

Eisen, J.A., D. Kaiser, and R.M. Myers. 1997. *Nature (Med.)* **3:** 1076–1078.

Eisen, J.A., K.S. Sweder, and P.C. Hanawalt. 1995. *Nucleic Acids Res.* **23:** 2715–2723.

Felsenstein, J. 1985. *Am. Nat.* **125:** 1–15.

Fitch, W.M. 1970. *Syst. Zool.* **19:** 99–113.

Gatesy, J., R. Desalle, and W. Wheller. 1993. *Mol. Phylog. Evol.* **2:** 152–157.

Goldman, N., J.L. Thorne, and D.T. Jones. 1996. *J. Mol. Biol.* **263:** 196–208.

Henikoff, S., E.A. Greene, S. Pietrovsky, P. Bork, T.K. Attwood, and L. Hood. 1997. *Science* **278:** 609–614.

Henikoff, S., S. Pietrokovski, and J.G. Henikoff. 1998. *Nucleic Acids Res.* **26:** 311–315.

Hillis, D.M. 1994. In *Homology: The hierarchical basis of comparative biology* (ed. B.K. Hall), pp. 339–368. Academic Press, San Diego, CA.

Maddison, W.P. and D.R. Maddison. 1992. *MacClade*. Sinauer Associates, Sunderland, MA.

Reeck, G.R., C. Haën, D.C. Teller, R.F. Doolittle, W.M. Fitch, R.E. Dickerson, P. Chambon, A.D. McLachlan, E. Margoliash, T.H. Jukes et al. 1987. *Cell* **50:** 667.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. In *Molecular systematics* (ed. D.M. Hillis, C. Moritz, and B.K. Mable), pp. 407–514. Sinauer Associates, Sunderland, MA.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. *Science* **278:** 631–637.

Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty et al. 1997. *Nature* **388:** 539–547.

Weisburg, W.G., J.G. Tully, D.L. Rose, J.P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T.G. Lawrence, J. Van Etten et al. 1989. *J. Bacteriol.* **171:** 6455–6467.

Yokoyama, S. 1997. *Annu. Rev. Genet.* **31:** 315–336.

Zardoya, R., E. Abouheif, and A. Meyer. 1996. *Trends Genet.* **2:** 496–497.